Analysis of Differential Item Functioning in Translated Assessment Instruments

Glen R. Budgell, Canadian Nurses Association
Nambury S. Raju, Illinois Institute of Technology
Douglas A. Quartetti, Georgia Institute of Technology

The usefulness of three IRT-based methods and the Mantel-Haenszel technique in evaluating the measurement equivalence of translated assessment instruments was investigated. A 15-item numerical test and an 18-item reasoning test that were originally developed in English and then translated to French were used. The analyses were based on four groups, each containing 1,000 examinees. Two groups of English-speaking examinees were administered the English version of the tests; the other two were French-speaking examinees who were administered the French version of the tests. The percent of items identified with significant differential item functioning (DIF) in this study was similar to findings in previous large-sample studies. The four DIF methods showed substantial consistency in identifying items with significant DIF when replicated. Suggestions for future research are provided. *Index terms: area measures, differential item functioning, item response theory, language translations, Lord's χ^2, Mantel-Haenszel procedure.*

When tests are adapted from the language and culture in which they were developed to another language, the measurement equivalence of the adapted instrument should be assessed. The original and adapted instruments may not be equivalent because (1) the meaning of the test items may have been inadvertently altered in the translation process and/or (2) the test items may not be equally relevant in the different linguistic and cultural groups. In the past, some cross-cultural researchers (Brislin, 1980, 1986) have argued that the conscientious implementation of standard translation procedures (i.e., back-translation and decentering) along with a variety of classical test theory (CTT) analyses (Gulliksen, 1950; Lord & Novick, 1968) of examinee responses will produce accurate linguistic translations that result in equivalent scales.

The use of CTT methods in these situations is problematic. Although CTT methods are considered appropriate for making within-group (or population) comparisons, they are inappropriate for making comparisons across groups (or populations) because of their population dependency. Consequently, although standard translation procedures and traditional statistical analyses may be necessary first steps, they are not sufficient to ensure measurement equivalence (Hulin, Drasgow, & Parsons, 1983). Therefore, different criteria and methods are needed to assess the quality of translated tests. Item response theory (IRT; Lord, 1980) is a framework that avoids the serious deficiencies inherent in CTT and therefore has several advantages in assessing measurement equivalence in translated tests.

Within the IRT framework, Drasgow has defined measurement equivalence as a property that exists "... when the relations between observed test scores and the latent attribute measured by the test are identical across subpopulations" (Drasgow, 1984, p. 134). For translated assessment instruments, tests are said to exhibit measurement equivalence when individuals who are equal in the trait measured by the test but who come from different cultural and linguistic groups have the same observed score. Unless equivalent assessment instruments are used, it is uncertain whether score differences represent group differences

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 19, No. 4, December 1995, pp. 309–321

0146-6216/95/040309-13$1.90

and similarities or measurement artifacts.

Statistical methods based on IRT make it possible to identify items in translated tests that do not function in the same way for the two linguistic groups under study. These items are said to display differential item functioning (DIF). Measurement equivalence can be achieved by identifying and eliminating DIF items from the test.

To date, only a relatively small number of studies have applied IRT-based DIF analysis to assess the measurement equivalence of translated assessment instruments (Candell & Hulin, 1987; Candell & Roznowski, 1984; de Vera, 1985; Drasgow & Hulin, 1989; Drasgow & Lissak, 1983; Ellis, 1989, 1991; Ellis, Minsel, & Becker, 1989; Hulin, Drasgow, & Komocar, 1982; Hulin & Mayer, 1986; Osberg, Scott, & Raju, 1985). Some of these IRT analyses were based on small samples and, in others, there was only limited control over the linguistic abilities of the samples used. Therefore, the major purpose of this study was to further investigate the measurement equivalence of translated assessment instruments with IRT-based DIF methods using adequate sample sizes and with better control over the linguistic ability of the samples. Because of their current popularity (Millsap & Everson, 1993), three different IRT-based methods were used in this study for assessing DIF: the signed area (SA) method (Raju, 1988, p. 496), unsigned area (UA) method (Raju, 1988, p. 496), and Lord's (1980) χ^2 (LC) method. The exact area methods (Raju, 1988, 1990) compare the area between the item response functions estimated in two different groups, whereas LC tests whether estimates of item parameters from the two groups are equal.

A secondary purpose was to examine the usefulness of the Mantel-Haenszel (1959; MH) technique in assessing the measurement equivalence of translated assessment instruments. The MH statistic is a non-IRT based method for assessing DIF (Dorans & Holland, 1993; Holland & Thayer, 1988; Raju, Bode, & Larsen, 1989), and it is very easy to use in practice. However, the usefulness of this technique in assessing the measurement equivalence of translated assessment instruments has not been studied. If MH-based DIF analysis could be used instead of IRT-based DIF analysis in assessing measurement equivalence in translated assessment instruments, then one of the practical disadvantages of IRT-based analysis—the need for large sample sizes—would be eliminated.

Method

Instruments

Two tests from a non-commercially available, professionally developed general cognitive ability test battery used in the federal Canadian Government were studied in this investigation (additional information about these tests may be obtained from the first author). The test battery is designed to assess examinee potential to perform basic administrative tasks such as planning, decision-making, analyzing and solving numerical problems, evaluating information (reasoning), and recognizing effective communication. The paper-and-pencil, group-administered test battery contains a total of 65 multiple-choice items and is available in both English and French. Only the 15-item Numerical Test and the 18-item Reasoning Test were used in this study.

The items were translated using a multistage translation process involving a team of professional translators and a committee of bilingual psychologists. This process was derived from integrating several of the traditional translation procedures described in Brislin (1980) and new methods to handle the unique problems associated with accurately translating psychological scales and items. All items were originally developed in English and then were translated into French by a team of translators. A committee of bilingual test development experts then reviewed the English and French versions of each item for linguistic equivalence. This committee was briefed on the purpose of the test and was provided with the test specifications, the ability each item was designed to assess and, when available, the reason for specific item distractors. The committee made revisions as required, and then both language versions of the test were pretested using an appropriate sample. Following the pretest, the committee of bilingual test development experts reviewed the English and French versions of each item taking the pretest item analysis data into consideration. The necessary linguistic revisions to the French items were then made.

Examinees

The available database contained 16,362 examinees who were administered one of the language versions of the test battery as part of an ongoing operational assessment program. The tests were administered between July, 1986 and March, 1991. Four samples containing examinees who met the following two conditions were created: (1) the examinee's native language matched the language version of the test he/she was administered, and (2) the examinee resided in a location in Canada in which the linguistic dominance of the environment matched the language version of the test he/she was administered. The purpose of these conditions was to provide some control over the linguistic proficiency and "cultural purity" of the examinees included in the samples.

The following four samples were generated by randomly selecting examinees without replacement from the total database so that the samples would be independent. English Sample 1 (E1) contained 1,000 examinees who had been administered the English version of the test and met the two criteria described above. English Sample 2 (E2) was similar to E1, also with 1,000 examinees. French Sample 1 (F1) contained 1,000 examinees who had been administered the French version of the test and met the above criteria. French Sample 2 (F2) was similar to F1, also with 1,000 examinees.

There were 286 males and 714 females in E1; 270 males and 730 females in E2. F1 contained 470 males and 530 females; F2 contained 491 males and 509 females. The average age of examinees in E1, E2, F1, and F2 was 34.1, 34.5, 32.5, and 32.1 years, respectively, and the corresponding average years of education were 14.3, 13.3, 12.6, and 14.9.

Sample Comparisons

Separately for both the Numerical Test and the Reasoning Test, the following four comparisons were made using the three IRT-based procedures (SA, UA, and LC) and the MH technique for identifying DIF: Comparison 1—E1 versus F1, Comparison 2—E2 versus F2, Comparison 3—E1 versus E2, and Comparison 4—F1 versus F2. To the extent that DIF in translated assessment instruments is measured consistently within and across the IRT and MH methods, approximately the same items should be identified as having significant DIF in Comparisons 1 and 2 and no items should be found to exhibit significant DIF in Comparisons 3 and 4. This pattern of results should be found for both the Numerical Test and the Reasoning Test.

Because different investigators may use different DIF methods in practice, agreement among the methods was examined separately for the Numerical and Reasoning Tests across all four comparisons. This type of convergence analysis has not been previously conducted in the context of translated assessment instruments. The total number of items identified with significant DIF was determined for each of the four DIF methods, and then the number of common items identified between methods was determined.

Because it was impossible to state which items were truly DIF items in this study, in order to maximize the probability of identifying only items with true DIF, the results from the MH method and the three IRT-based DIF analyses were examined by combining the results of Comparisons 1 and 2. In this analysis, only those items that displayed significant DIF in both Comparisons 1 and 2 were identified as items with significant DIF. Agreement among the four DIF methods was evaluated using these replicated results.

Computation of DIF: IRT Procedures

Unidimensionality. The English and French versions of both the Numerical and Reasoning Tests were separately factor analyzed, using principal components analysis (PCA), to determine the degree to which the items formed unidimensional sets. Reckase's (1979) eigenvalue criterion for assessing unidimensionality was used. The English versions of the two tests were analyzed using the E1 sample and the French versions of the two tests were analyzed using the F1 sample.

Estimation of item parameters. The 15-item Numerical Test and the 18-item Reasoning Test were calibrated separately for each test and sample (E1, E2, F1, and F2) using BILOG 3 (Mislevy & Bock, 1990). Because of the estimation problems associated with

the *c* parameter in the three-parameter logistic model (Lord, 1980), the two-parameter logistic model (2PLM; Lord, 1980, pp. 12–13) was used. The default option in BILOG 3, marginal maximum a posteriori (MMAP) estimation, was used in each of the eight calibrations. Mislevy and Bock indicate that no dependable, formal test currently exists to assess the goodness of fit of the IRT model to the data for short tests (11 to 20 items). However, they suggest that useful information about the fit of individual items may be obtained by inspecting standardized posterior residuals.

Linking of item parameter estimates. The item parameter estimates were transformed as follows for both tests across each of the four comparisons: Comparison 1—estimates from the F1 sample were transformed to the scale underlying the E1 sample; Comparison 2—estimates from F2 were transformed to the E2 scale; Comparison 3—estimates from E2 were transformed to the E1 scale; Comparison 4—estimates from F2 were transformed to the F1 scale.

Because IRT-based DIF analysis requires that the estimated item parameters from the two subpopulations [commonly referred to as the reference and focal groups (Holland & Wainer, 1993)] be put on a common scale prior to any DIF analysis, the test characteristic curve procedure of Stocking & Lord (1983), as implemented in the EQUATE computer program (Baker, 1993), was used. In order to minimize the effects of biased items on item parameter rescaling and DIF detection, an iterative application of EQUATE was used. This iterative procedure, described in Candell & Drasgow (1988), is computationally less intensive than the noniterative "purification" approach of Lord (1980) and appears to lead to more accurate detection of DIF (Candell & Drasgow, 1988; Kim & Cohen, 1992; Park & Lautenschlager, 1990).

Statistical tests for DIF indexes. Using the transformed item parameter estimates obtained in the previous step for each comparison and separately for each test, SA and UA were computed using Raju's (1988) equations. Raju's (1990) z statistics for SA and UA were computed to identify items with significant DIF. For the z statistics associated with SA and UA, a two-tailed test of $z < -2.58$ or $z > +2.58$ ($\alpha = .01$) was used to identify items with significant DIF.

LC for the equality of the *a* and *b* parameters was computed for each item using the transformed item parameter estimates (Hulin et al., 1983). A χ^2 of 9.21 [$\alpha = .01$ with 2 degrees of freedom (df)] was used to identify items with significant DIF.

Computation of DIF: MH procedure

As with other χ^2 procedures, MH compares the performance of two subpopulations on each item across different score groups. Similar to the IRT-based analysis, MH χ^2 statistics were computed for each item across each of the four comparisons to identify which items had significant DIF. A total of eight MH-based DIF analyses, with no iterative purification, were conducted. The item under consideration was included in forming the score groups, and each possible total number-correct score was used to define a separate score group. A χ^2 of 6.63 ($\alpha = .01$, 1 df) was used to identify items with significant DIF.

Results

The number-correct score summary statistics for E1, E2, F1, and F2 on the Numerical and Reasoning Tests are shown in Table 1. The number-correct score means, standard deviations (SDs), and the Kuder-Richardson Formula 20 (KR20) estimates of reliability on the Numerical Test were approximately the same across the four samples. The number-correct score means, SDs, and KR20s on the Reasoning Test were also similar across the four samples.

Table 1
Mean, Standard Deviation (SD), and KR20 for Two English and Two French Samples on the Numerical and Reasoning Tests

	Numerical Test			Reasoning Test		
Sample	Mean	SD	KR20	Mean	SD	KR20
E1	6.87	2.57	.56	9.29	2.87	.59
E2	6.81	2.57	.56	9.27	2.80	.56
F1	6.97	2.61	.58	8.48	2.69	.50
F2	7.13	2.56	.56	8.58	2.77	.53

Unidimensionality

For E1, the PCA of the English version of the Numerical Test yielded six eigenvalues larger than

1.00, with the largest eigenvalue of 3.34 accounting for 22.3% of the total variance. Analysis of the English version of the Reasoning Test yielded seven eigenvalues greater than 1.00, with the largest eigenvalue of 3.50 accounting for 19.5% of the total variance.

For F1, the PCA of the French version of the Numerical test yielded three eigenvalues greater than 1.00, with the largest eigenvalue of 3.54 accounting for 23.6% of the total variance. Analysis of the French version of the Reasoning Test yielded six eigenvalues greater than 1.00, with the largest eigenvalue of 3.03 accounting for 16.8% of the total variance.

The PCAs, using interitem tetrachoric correlations, yielded similar structures for both the English and French versions of the Numerical and Reasoning Tests. The percent of total variance associated with the largest eigenvalue for both language versions of the Reasoning Test did not quite meet Reckase's minimum criterion of 20% for unidimensionality (Reckase, 1979), but a scree test (Tatsuoka, 1988) suggested that there was a single dominant factor in both tests. Nevertheless, there is evidence that certain IRT models can be applied to moderately heterogeneous item sets (Drasgow & Parsons, 1983).

IRT Calibrations

Because the Numerical and Reasoning Tests contained less than 20 items, the standardized posterior residuals rather than the χ^2s (for goodness of fit) generated by BILOG 3 were reviewed for determining the overall fit of the items. Following the recommendation of Mislevy & Bock (1990, pp. 1–10, 4–27), residual values greater than 2.0 along with their posterior weights were taken to indicate lack of fit between the 2PLM and the data.

For the Numerical Test, only Item 8 did not fit the 2PLM in both French samples, and Items 8, 10, and 15 did not fit the 2PLM in both English samples. On the Reasoning Test, only Item 3 did not fit the 2PLM in both French samples and Item 10 did not fit the 2PLM in both English samples. The overall fit of the items in the Numerical and Reasoning Tests to the 2PLM was considered adequate across the two English and two French samples and, therefore, all items in both tests were included in subsequent analyses for DIF.

DIF Indexes for the Numerical Test

The DIF statistics for Comparisons 1 and 2 are provided in Table 2. For each item, Table 2 shows the SA and UA and the associated z statistics, LC, and the MH statistics (α and χ^2). The expected MH α for an item with no DIF is 1.0; significant deviations from this expected value will typically lead to significant MH χ^2 values. In each of the four comparisons, items with significant DIF are identified with an asterisk. Results from Comparisons 3 and 4 are not presented here because there were no items with significant DIF.

SA identified four items with significant DIF in Comparison 1 and five items with significant DIF in Comparison 2 (Table 2). Four of these items (Items 5 and 7–9) were common to both comparisons. UA identified four and five items with significant DIF in Comparisons 1 and 2, respectively, and three items (Items 7–9) were common to both comparisons. Similarly, LC identified six items with significant DIF in each comparison, with five items (Items 5, 7–9, and 13) common to both comparisons. MH identified seven items with significant DIF in Comparison 1 and six items with significant DIF in Comparison 2. Five of these items (Items 5, 7–9, and 13) were common to both comparisons. Finally, in Comparison 1, 27%, 27%, 40%, and 47% of numerical items had significant DIF, respectively, for SA, UA, LC, and MH; respective results for Comparison 2 were 33%, 33%, 40%, and 40%.

Table 3 shows the degree of overlap between the four DIF procedures. Note that in this overlap analysis, only items that showed significant DIF in both comparisons were considered in order to optimize the probability of including only "truly" biased items. The numbers in the diagonal refer to the number of items identified as biased by a DIF method across both comparisons. UA identified the smallest number of items (3) as having significant DIF. The same three items (Items 7–9) also were identified as having significant DIF by the other three methods. LC and MH identified the largest number of items (5) as having significant DIF; also, the same

Table 2
DIF Indexes for Comparisons 1 and 2 on the Numerical Test

Comparison and Item	SA	z	UA	z	LC	MH α	MH χ^2
Comparison 1 (E1-F1)							
1	−.21	−.22	.21	−.22	.15	.88	.82
2	−.67	−1.57	.67	−1.52	3.53	.58	12.55*
3	−.32	−1.69	.46	1.60	1.35	.85	2.14
4	.19	1.14	.31	1.06	2.34	1.04	.11
5	−1.33	−4.41*	1.33	−4.42*	20.06*	.53	44.08*
6	.92	2.45	1.16	−1.91	8.69*	1.33	7.10*
7	1.22	7.12*	1.23	−6.88*	59.33*	2.59	87.76*
8	5.00	3.74*	5.00	3.74*	17.31*	1.95	52.31*
9	−1.12	−6.44*	1.24	4.74*	34.48*	.38	88.72*
10	−.55	−1.76	1.37	1.78	3.02	.89	1.26
11	.22	1.30	.58	−1.94	6.17	.97	.04
12	−.40	−1.97	.94	1.90	3.21	.83	3.25
13	.28	2.09	.28	2.09	15.79*	1.42	9.02*
14	.01	.10	.28	1.40	5.28	1.07	.38
15	.86	.96	1.02	−.93	1.64	1.06	.14
Comparison 2 (E2-F2)							
1	−.60	−.51	.47	−.42	.95	1.11	.50
2	−.05	−.08	.43	.73	3.35	.70	5.34
3	−.17	−.94	.14	−1.04	4.61	.69	10.85*
4	.14	.81	.36	1.52	1.52	1.06	.29
5	−1.09	−4.08*	1.38	−2.14	22.27*	.54	39.04*
6	−.11	−.26	.41	.58	.21	.97	.07
7	1.04	6.93*	1.14	−7.39*	51.14*	2.29	64.02*
8	7.83	4.40*	7.58	4.51*	19.84*	2.47	94.91*
9	−1.27	−6.69*	1.10	−5.99*	44.56*	.35	105.77*
10	−.32	−.80	.30	.43	1.57	.84	2.80
11	.17	1.07	.25	2.40	2.45	1.09	.55
12	−.38	−1.69	.67	1.27	2.79	.81	4.58
13	.48	3.61*	.54	−3.89*	14.08*	1.37	7.40*
14	.10	.68	.71	2.98*	13.80*	1.25	4.07
15	1.52	1.90	1.51	−1.65	3.26	1.13	.89

*Statistically significant at the .01 level.

five items (Items 5, 7–9, and 13) were identified by both procedures. SA identified one more item (Item 5) as having significant DIF than did UA.

Table 3
Number and Overlap of Common DIF Items Across Methods in the Numerical Test

Method	SA	UA	LC	MH
SA	4*	3	3	3
UA		3*	3	3
LC			5*	5
MH				5*

*Number of items identified as biased in Comparisons 1 and 2 (see Table 2).

DIF Indexes for the Reasoning Test

The DIF statistics for the Reasoning Test are shown in Table 4 for Comparisons 1 and 2. Data from Comparisons 3 and 4 for the Reasoning Test are not provided because no items were identified as having significant DIF.

SA identified eight items with significant DIF in Comparison 1 and six items with significant DIF in Comparison 2 (Table 4). Six of the items (Items 2, 7, 8, 12, 15, and 18) were common to both comparisons. UA identified ten items with significant DIF in Comparison 1 and two items with significant DIF in

Table 4
DIF Indexes for Comparisons 1 and 2 on the Reasoning Test

Comparison and Item	SA	z	UA	z	LC	MH α	MH χ^2
Comparison 1 (E1–F1)							
1	.71	.98	.71	−1.01	4.64	1.54	9.74*
2	−4.60	−3.70*	4.88	3.32*	13.11*	.35	99.52*
3	.79	.35	1.14	.41	3.39	1.28	4.46
4	1.10	1.52	2.17	1.13	5.70	1.37	10.76*
5	−.87	−1.94	2.82	−1.63	6.56	1.08	.52
6	.30	.38	.25	.36	10.13*	1.72	19.70*
7	−1.92	−6.45*	2.02	6.90*	32.94*	.50	50.18*
8	−.67	−4.10*	.80	5.02*	.55	.96	.09
9	−.83	−4.29*	.89	4.81*	2.41	.86	2.16
10	−3.37	−1.19	3.85	1.39	19.92*	.46	30.36*
11	1.14	1.78	1.18	1.56	14.77*	2.39	56.89*
12	−1.15	−6.52*	1.23	−6.45*	12.06*	.70	12.84*
13	.40	.91	.66	−1.40	6.70*	1.55	14.45*
14	−.23	−1.26	.86	3.27*	5.28	1.10	.74
15	−.84	−6.51*	.96	6.77*	4.86	.87	2.02
16	−.48	−3.82*	.60	8.97*	5.33	.96	.10
17	−.25	−2.09	.80	3.99*	13.02*	1.59	21.06*
18	−.53	−5.99*	.60	7.23*	1.07	1.15	1.41
Comparison 2 (E2–F2)							
1	.92	1.37	.95	−1.40	6.16	1.80	19.16*
2	−3.32	−2.91*	3.20	2.82*	21.27*	.43	70.31*
3	−2.53	−.88	4.44	−1.20	3.79	1.51	13.05*
4	1.79	2.02	4.67	2.20	9.17*	1.25	5.44
5	−.29	−.78	1.50	−1.25	5.11	1.14	1.79
6	.44	.53	.95	1.81	34.31*	2.05	34.82*
7	−1.90	−4.10*	1.79	2.70*	19.93*	.58	32.07*
8	−.52	−3.25*	.35	−2.26	1.82	.85	2.28
9	−.44	−1.88	.53	.87	.95	.89	1.32
10	−3.47	−1.35	3.52	1.39	13.28*	.38	45.72*
11	.51	1.10	.77	1.64	12.17*	1.71	21.77*
12	−.80	−4.65*	1.05	−2.20	9.57*	.76	7.33*
13	.01	.04	.11	.35	2.53	1.25	3.85
14	−.24	−1.20	.11	.33	.55	1.11	.98
15	−.63	−3.70*	.50	−2.25	4.86	.82	4.03
16	−.16	−1.01	.21	.97	1.61	1.25	3.82
17	−.25	−2.15	.33	−1.40	3.05	1.19	2.79
18	−.36	−3.78*	.14	1.87	.12	1.17	1.69

*Statistically significant at the .01 level.

Comparison 2; two of the items (Items 2 and 7) were common to both comparisons. LC identified eight (Comparison 1) and seven (Comparison 2) items with significant DIF, with six items (Items 2, 6, 7, and 10–12) common to both comparisons. Finally, MH identified ten and eight items with significant DIF in Comparisons 1 and 2, respectively; seven of the items (Items 1, 2, 6, 7, and 10–12) were common to both comparisons.

Table 5 shows the amount of overlap between the four DIF procedures. As in the case of the Numerical Test, only items that showed significant DIF in both comparisons were included in this analysis. UA identified the fewest items (2) as biased, whereas MH identified the most items (7) as biased. The two items (Items 2 and 7) that were identified as biased

by UA also were identified as biased by the other three DIF techniques. All six items identified as biased by LC also were identified as biased by MH. SA identified six items (Items 2, 7, 8, 12, 15, and 18) as having significant DIF, of which only three items (Items 2, 7, and 12) were common to MH and LC.

Table 5
Number and Overlap of Common DIF Items Across Methods in the Reasoning Test

Method	SA	UA	LC	MH
SA	6*	2	3	3
UA		2*	2	2
LC			6*	6
MH				7*

*Number of items identified as biased in Comparisons 1 and 2 (see Table 4).

Discussion

Consistency in Detecting DIF

Consistency in detecting significant DIF across samples is necessary for a method to provide useful information. It was expected that in Comparisons 1 and 2 approximately the same items would be identified with significant DIF within each of the statistical methods, and that no items exhibiting significant DIF would be found in Comparisons 3 and 4. Data in Table 2 for the Numerical Test and in Table 4 for the Reasoning Test show that the overlap in items with significant DIF in Comparisons 1 and 2 was relatively high for each method, except for the UA analysis of the Reasoning Test in which the overlap was low. As expected, none of the methods revealed significant DIF in Comparisons 3 and 4.

Further examination of the UA analysis of the Reasoning Test showed that 10 items were identified with significant DIF in Comparison 1, whereas only two items were detected in Comparison 2. The two items detected in Comparison 2 were also detected in Comparison 1. It was expected that the UA method would produce results similar to those obtained with LC; therefore, it was surprising that the largest difference occurred using this method. In previous studies (Cohen & Kim, 1993; Raju, Drasgow, & Slinde, 1993), UA and LC produced similar results.

Correspondence Between Methods

When several DIF methods are available for the detection of measurement nonequivalence at the item level, it is useful to know the degree to which the different methods yield similar results. Therefore, correspondence between the different methods was examined by determining the number and overlap of items identified with significant DIF in Comparisons 1 and 2, separately for the Numerical and Reasoning Tests. The four methods identified between four and seven items with significant DIF on the Numerical Test and between two and ten items on the Reasoning Test. In general, the MH procedure detected the most items with significant DIF, followed by LC, and then followed by SA and UA.

With respect to the area methods, Raju (1988) indicated that SA reflects only the difference in the *b*s (in the 2PLM), whereas UA reflects the differences in both *a* and *b*. Consequently, it should be expected that items with significant SA would also have significant UA and, in general, there should be at least as many items with significant UA as there are with significant SA. In this study, the expected pattern was evident in three of the four analyses; however, in Comparison 2 on the Reasoning Test, SA detected six items with significant DIF, whereas UA detected only two items.

Swaminathan & Rogers (1990) found that MH identified uniform DIF quite well, but it was less successful in detecting nonuniform DIF. In other words, items that differed in the *a* parameter were not consistently identified as biased by MH. This was not the case here. MH consistently identified more items with significant DIF on both the Numerical and Reasoning Tests than any of the three IRT-based methods. According to Millsap & Everson (1993), the MH technique may have an inflated α problem, especially when the total number of items in a test is less than 20.

The agreement between MH and LC was substantial. In the Numerical Test, both methods identified the same five items as having significant DIF (Table 3); in the Reasoning Test, six items that were identified as biased with LC were also identified as biased with MH. At least in this application (fidelity

of language translations), MH and LC appear to identify mostly the same items as biased.

Item Review by the Translation Review Committee

A follow-up analysis of DIF was undertaken by a review committee. The purpose of this special analysis was to assess the degree to which the review committee agreed with the statistical techniques in identifying DIF. This type of review is essential for diagnosing the underlying causes of DIF and hopefully controlling for them in future translations of assessment instruments.

A review committee comprised of two bilingual psychologists and an expert translator was established to review and rate the translation quality of 13 of the 33 items evaluated in this study (6 of the 15 items from the Numerical Test and 7 of the 18 items from the Reasoning Test). Each member of the translation review committee independently compared the English version of an item to the French version and rated the quality of the translation using a five-point scale: 1 = *major problems*, 2 = *some major problems*, 3 = *minor problems*, 4 = *acceptable*, and 5 = *fully acceptable*. In rating the items, the committee members examined carefully the linguistic equivalence and the cultural appropriateness of words and expressions to ensure equivalence. Three items with the most similar item response functions (two items from the Numerical Test and one item from the Reasoning Test) and ten items with the greatest DIF indexes in one or both comparisons (four items from the Numerical Test and six items from the Reasoning Rest) were selected for this review. The items were arranged in the same random order for each committee member.

Results of this review are provided in Table 6. The translator (Reviewer 3) rated all items as *acceptable* (4) or *fully acceptable* (5). The bilingual psychologists rated as *acceptable* or *fully acceptable* the three items that displayed measurement equivalence (i.e., no significant DIF detected—Items 1 and 15 on the Numerical Test and Item 14 on the Reasoning Test). Of the 10 items included in the review that were identified as having significant DIF, one bilingual psychologist (Reviewer 1) rated five of these items as having *minor* to *major problems*, whereas the other bilingual psychologist (Reviewer 2) rated four items as having *minor* to *major problems*. Reviewers 1 and 2 (the bilingual psychologists) gave their lowest ratings to the same three items (Item 8 from the Numerical Test and Items 1 and 2 from the Reasoning Test).

The fact that the statistical procedures identified many more items as problematic than did the review committee cannot be taken as evidence against the usefulness of the statistical procedures. It seems more likely that traditional expert analysis may not

Table 6
Results of Item Review by the Translation Review Committee

Numerical Test Item No.	Reasoning Test Item No.	Significant DIF Detected	Reviewer 1	2	3
1		NO	5	5	5
15		NO	5	4	5
	14	NO	5	4	5
5		YES (All Methods)	3	4	5
7		YES (All Methods)	4	5	5
8		YES (All Methods)	2	2	4
9		YES (All Methods)	5	5	5
	1	YES (MH Only)	2	1	4
	2	YES (All Methods)	2	2	4
	3	YES (MH Only)	5	5	5
	7	YES (All Methods)	5	3	5
	15	YES (SA and UA Only)	5	5	4
	18	YES (SA Only)	3	5	5

be sufficient to detect all the nuances and subtleties of language that could contribute to the measurement nonequivalence of translated tests. Results of this special analysis also suggest that successful translation may require the use of individuals who not only are bilingual but who have knowledge in test and item development.

Comparison to Previous Studies

Table 7 provides a summary of results found in this study and previous studies applying IRT-based procedures for identifying DIF in translated assessment instruments. These studies all used unilingual samples. Table 7 shows that across the previous studies the number of items identified with significant DIF ranged from a low of 1.5% to a high of 64%. Because Lord's IRT-based procedure was used to detect DIF in most of the previous studies, only results from LC were used in this analysis. In the current study, 40% of the items on the Numerical Test were identified with significant DIF in both Comparisons 1 and 2; on the Reasoning Test, 44% were identified in Comparison 1 and 39% in Comparison 2. Although the previous studies differ from each other and with the current study in many respects, given the comprehensive translation process used in this study and the control used in selecting the samples, it was expected that the percent of items identified as having significant DIF in this study would be lower than in previous studies, which tended to be somewhat less rigorous in these respects. The percent of items identified as having significant DIF in this study tended to be in the midrange.

There are several possible reasons why more items than expected were identified as having significant DIF in this study. First, as indicated earlier, the unidimensionality assumption and the fit of the data to the IRT model were only marginally met. As shown in Table 1, the test reliabilities were low. These conditions could have had an effect on the accurate estimation of item parameters and the subsequent DIF analysis. Second, the item types contained in both of the tests tended to be longer and contained more information than most standard item types. Given the complex nature of the items, there may have been a greater chance for translation errors to occur. Third, a review of the previous studies also suggests that the identification of items with significant DIF may be related to sample size. Inspection of Table 7 shows that in studies in which the sample sizes were relatively small (e.g., Ellis, 1989, 1991), very few items were identified as having significant DIF, whereas in studies that were based on large sample sizes such as Drasgow & Lissak (1983), Osberg et al. (1985), and the current study, a much higher percent of items were identified with significant DIF.

Conclusions and Limitations

The results contribute to the increased understanding of the usefulness of IRT and MH-based DIF analysis in evaluating the measurement equivalence of translated assessment instruments. Although many of the analyses in this study provided support for the usefulness of these statistical methods in detecting DIF in translated assessment instruments, some did not. These mixed findings may be due to the following limitations.

First, the PCA of the French and English versions of the Numerical and Reasoning Tests used in this study indicated that the tests did not quite meet Reckase's criterion for unidimensionality. An assumption of the 2PLM used in this study is that the tests are unidimensional. Although Drasgow & Parsons (1983) reported that the 2PLM can be applied to moderately heterogeneous item sets, violation of this assumption could have had an effect on the accurate estimation of the item parameters and the subsequent DIF analysis.

Second, the fit of the data to the IRT model was only marginally met. This problem may be related to the first issue regarding the moderate heterogeneity of the tests used here. The marginal fit of the data to the IRT model could also have had an effect on the accurate estimation of item parameters and the subsequent DIF analysis. In the future, a multidimensional IRT model (McKinley & Reckase, 1983) could be applied to the data to determine if a better fit between the data and the model could be obtained and to determine the impact this would have on the DIF analyses.

Table 7
Summary of Previous Studies Using IRT-Based Procedures to Assess Measurement Equivalence in Translated Tests

Study	Instrument	Sample	Results
Drasgow & Lissak (1982)	Otis-Lennon Ability Test (Translated from English to French)	French: 700 English: 1,400	50 Items DIF: 32 (64%)
de Vera (1985)	Job Description Index (Translated from English to Filipino)	Filipino: 603 English: 500	33 Items DIF: 11 (33.3%)
Osberg, Scott, & Raju (1985)	Reading Test (Translated from English to Spanish)	Spanish: 785 English: 1,252	16 Items DIF: 10 (62.5%)
Hulin & Mayer (1986)	Job Description Index (Translated from English to Hebrew)	Hebrew: 500 English: 308	66 Items DIF: 21 (31.8%)
Candell & Hulin (1987)	Job Description Index (Translated from English to French)	French: 349 English: 435	86 Items DIF: 15 (17.4%)
Drasgow & Hulin (1989)	Job Description Index (Translated from English to Spanish)	Spanish: 346 English: 486	68 Items DIF: 21 (30.9%)
Ellis (1989)	Career Ability Test (Translated from English to German)	German: 205	106 Items DIF: 8 (7.6%)
	WILDE-Intelligence Test (Translated from German to English)	English: 217	145 Items DIF: 2 (1.4%)
Ellis (1991)	Attitude Survey (CCMHS) (Translated from German to English)	German: 205 English: 197	151 Items DIF: 13 (8.6%)
	Trierer Persönlichkeitsfragebogen (TPF: Becker, 1989) (Translated from German to English)	German: 213 English: 295	120 Items DIF: 11 (9.2%)
Current Study	Numerical Ability Test (Translated from English to French)	English: 1,000 French: 1,000	15 Items DIF: 6 (40%)[a] DIF: 6 (40%)[b]
	Reasoning Ability Test (Translated from English to French)	English: 1,000 French: 1,000	18 Items DIF: 8 (44%)[a] DIF: 7 (39%)[b]

[a]Based on LC data from Comparison 1.
[b]Based on LC data from Comparison 2.

Third, the Numerical Test contained only 15 items, and the Reasoning Test contained 18 items. The relatively small number of items per test affected the KR20 estimates of reliability. The use of the number-correct score with low reliability as the matching variable in forming score groups also may have contributed to DIF identification with the MH technique.

Fourth, the statistical methods may only be useful in accurately detecting DIF in certain types of translated items. Hulin & Mayer (1986), Candell & Hulin (1987), and Drasgow & Hulin (1989) have reported success in applying IRT in identifying DIF in an attitude survey (i.e., the Job Description Index) that has been translated into three different languages. Drasgow & Lissak (1983), Osberg et al. (1985), Ellis (1989), and the present study used various types of ability items; the percent of items detected as having significant DIF across these studies varied widely. Future research should examine whether these statistical methods can be applied to instruments measuring all types of psychological variables.

References

Baker, F. B. (1993). *EQUATE2: Computer program for equating two metrics in item response theory* [Com-

puter program]. Madison: University of Wisconsin, Laboratory of Experimental Design.

Brislin, R. (1980). Translation and content analysis of oral and written material. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology* (Vol. 2; pp. 389–444). Boston: Allyn and Bacon.

Brislin, R. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137–164). Beverly Hills: Sage.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12,* 253–260.

Candell, G. L., & Hulin, C. L. (1987). Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence. *Journal of Cross-Cultural Psychology, 17,* 417–440.

Candell, G. L., & Roznowski, M. (1984, August). *Using IRT to establish equivalence across U.S. and Canadian subpopulations.* Paper presented at the annual meeting of the American Psychological Association, Toronto.

Cohen, A. S., & Kim, S.-H. (1993). A comparison of Lord's χ^2 and Raju's area measures on detection of DIF. *Applied Psychological Measurement, 17,* 39–52.

de Vera, M. V. (1985). Establishing cultural relevance and measurement equivalence using Emic and Etic items. *Dissertation Abstracts International, 46-07B,* 2485.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale NJ: Erlbaum.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95,* 134–135.

Drasgow, F., & Hulin, C. L. (1989). *Cross-cultural measurement.* Unpublished manuscript, University of Illinois at Urbana-Champaign.

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously-scored item responses. *Journal of Applied Psychology, 68,* 363–373.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7,* 189–199.

Ellis, B. (1989). Differential item functioning: Implications for test translators. *Journal of Applied Psychology, 74,* 912–921.

Ellis, B. (1991). Item response theory: A tool for assessing the equivalence of translated tests. *International Test Bulletin, 32,* 33–51.

Ellis, B. B., Minsel, B., & Becker, P. (1989). Evaluation of attitude survey translations: An investigation using item response theory. *International Journal of Psychology, 24,* 661–684.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel Procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale NJ: Erlbaum.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale NJ: Erlbaum.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Application of item response theory to analysis of attitude translations. *Journal of Applied Psychology, 67,* 818–825.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood IL: Dow Jones-Irwin.

Hulin, C. L., & Mayer, L. M. (1986). Psychometric equivalence of a translation of the JDI into Hebrew. *Journal of Applied Psychology, 71,* 83–94.

Kim, S.-H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29,* 51–66.

Lord. F. (1980). *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719–748.

McKinley, R., & Reckase, M. (1983, April). *The use of IRT analysis on dichotomous data from multidimensional tests.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Millsap, R. J., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

Mislevy, R., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models.* Mooresville IN: Scientific Software.

Osberg, D. W., Scott, J. C., & Raju, N. S. (1985, April). *An analysis of the use of item response theory to investigate the fidelity of test translations.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Park, D. G., & Lautenschlager, G. T. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement, 14,* 163–173.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495–502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14,* 197–207.

Raju, N. S., Bode, R. K., & Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. *Applied Measurement in Education, 2,* 1–13.

Raju, N. S., Drasgow, F., & Slinde, J. A. (1993). An empirical comparison of the area methods, Lord's chi-square test, and the Mantel-Haenszel technique for assessing differential item functioning. *Educational and Psychological Measurement, 53,* 301–314.

Reckase, M. D. (1979). Unifactor latent trait models applied to multi-factor tests: Results and implications. *Journal of Educational Statistics, 4,* 207–230.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361–370.

Tatsuoka, M. M. (1988). *Multivariate analysis: Techniques for educational and psychological research.* New York: Macmillan.

Acknowledgments

The authors express their appreciation to the editor and two anonymous reviewers for their many helpful suggestions.

Author's Address

Send requests for reprints or further information to Glen R. Budgell, Canadian Nurses Association, 50th Driveway, Ottawa, Ontario, Canada K2P 1E2 or Nambury S. Raju, Institute of Psychology, Illinois Institute of Technology, Chicago IL 60616, U.S.A.